Leveraging spiking deep neural networks to understand neural mechanisms underlying selective attention

Lynn K. A. Sörensen[1,2*], Davide Zambrano[3,4], Heleen A. Slagter[5,6], Sander M. Bohté[3], H. Steven Scholte[1,2]

[1]Department of Psychology, University of Amsterdam, The Netherlands
[2]Amsterdam Brain & Cognition (ABC), University of Amsterdam, The Netherlands
[3]Machine Learning Group, Centrum Wiskunde & Informatica, The Netherlands
[4]Laboratory of Intelligent Systems, École Polytechnique Fédérale de Lausanne, Switzerland
[5] Department of Experimental and Applied Psychology, Vrije Universiteit Amsterdam, The Netherlands
[6] Institute of Brain and Behaviour Amsterdam, Vrije Universiteit Amsterdam, The Netherlands

* lynn.soerensen@gmail.com

Abstract

Spatial attention enhances sensory processing of goal-relevant information and improves perceptual sensitivity. The specific mechanisms linking neural changes to changes in performance are still contested. Here, we examine different attention mechanisms in spiking deep convolutional neural networks. We directly contrast effects of noise suppression (precision) and two different gain modulation mechanisms on performance on a visual search task with complex real-world images. Unlike standard artificial neurons, biological neurons have saturating activation functions, permitting implementation of attentional gain as gain on a neuron's input or on its outgoing connection. We show that modulating the connection is most effective in selectively enhancing information processing by redistributing spiking activity, and by introducing additional task-relevant information, as shown by representational similarity analyses. Precision did not produce attentional effects in performance. Our results, which mirror empirical findings, show that it is possible to adjudicate between attention mechanisms using more biologically realistic models and natural stimuli.

Introduction

Spatial attention is crucial for enabling naturalistic behaviour, such as crossing the street, where one needs to dynamically prioritize processing of goal-relevant aspects of the environment. Spatial attention is generally thought to increase the signal-to-noise ratio of activity in sensory regions representing the attended location. Yet, currently, there are several proposals on how this could be implemented. Some theories propose that attention selectively amplifies the neural signal by changing a neuron's gain (e.g. Martinez-Trujillo and Treue, 2004; Reynolds and Heeger, 2009), while others implement attentional selection through increased reliability or precision in processing, thereby emphasizing noise reduction rather than signal amplification (Feldman and Friston,

2010; Parr and Friston, 2017; Yu and Dayan, 2005). Computational models may allow for arbitration between these different ideas, yet existing models either examined attentional mechanisms in simplified sensory conditions (Feldman and Friston, 2010; Yu and Dayan, 2005) or were not developed to predict changes in performance (e.g. Beuth and Hamker, 2015; Ma et al., 2006; Reynolds and Heeger, 2009; Rothenstein and Tsotsos, 2014). As a result, it remains unclear how spatial attention may facilitate the processing of task-relevant information and thereby performance in more naturalistic settings: through gain modulation, noise suppression (precision) or a combination of both.

Deep convolutional neural networks (DCNN) are a way to close this gap in knowledge by linking changes in processing to performance in a fully controlled yet statistically rich setting (Kietzmann et al., 2019a; Richards et al., 2019; Scholte, 2018; Yamins and DiCarlo, 2016). Intriguingly, these networks not only parallel human performance on some object recognition tasks (VanRullen, 2017), but they also feature processing characteristics that bear a lot of resemblance to the visual ventral stream in primates (Eickenberg et al., 2017; Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Kubilius et al., 2018; Schrimpf et al., 2020; Yamins et al., 2014) . Leveraging this link between neural processing and performance has already enhanced insight into the potential mechanisms underlying shape perception (Kubilius et al., 2016), scene segmentation (Seijdel et al., 2020) and the role of recurrence during object recognition (Kar et al., 2019; Kietzmann et al., 2019b). DCNNs may thus provide a promising avenue for systematically investigating how different attention mechanisms may modulate neural processing and thereby, performance.

Here we use a recently developed class of networks, spiking deep convolutional neural networks (sDCNNs, Zambrano et al., 2018), that combine state-of-the-art performance with biologically inspired processing to arbitrate between different proposals of how selective attention may be neurally implemented. Findings from recent studies using DCNNs suggest that changing a neuron's gain is a viable way to implement selective processing in DCNNs (Lindsay and Miller, 2018; Luo et al., 2020, for a review, see Lindsay, 2020). Yet, these studies did not directly contrast different possible attention mechanisms. Spiking DCNNs as used here provide important additional constraints that can be used to evaluate the feasibility of different mechanisms. Due to their integrated neuron models, information is passed through these networks in temporal spike trains. This makes it possible to measure firing rates, examine population information, and estimate neural latencies and detection times in the network's output, and then to compare the effects on these measures of different manipulations in relation to findings from neuroscientific studies of attention.

For studying mechanisms of selective attention, sDCNNs provide two additional advantages compared to DCNNs: First, due to the neuron models that replace the activation functions used in DCNNs, sDCNN have a more realistic activation regime throughout the network. Commonly, DCNNs use a rectified linear units (ReLU, Nair and Hinton, 2010), while sensory neurons feature an activation that is more sigmoidal-like (e.g. Dayan and Abbott, 2001) and that saturates at high values (Naka and Rushton, 1966). This distinction becomes important when applying multiplicative gain: While for ReLUs there is no difference between the modulation of the in- or output, for sigmoidal activation functions there is a marked asymmetry, which either leads to input or response gain profiles (Ayaz and Chance, 2009; Martínez-Trujillo and Treue, 2002; Reynolds and Heeger, 2009). For understanding how selective attention may modulate the gain of neurons, this is a crucial feature that will also help to further clarify the implications of earlier DCNN studies for neural processing (Lindsay and Miller, 2018; Luo et al., 2020).

Secondly, sDCNNs feature task-unrelated noise due to their signal transmission

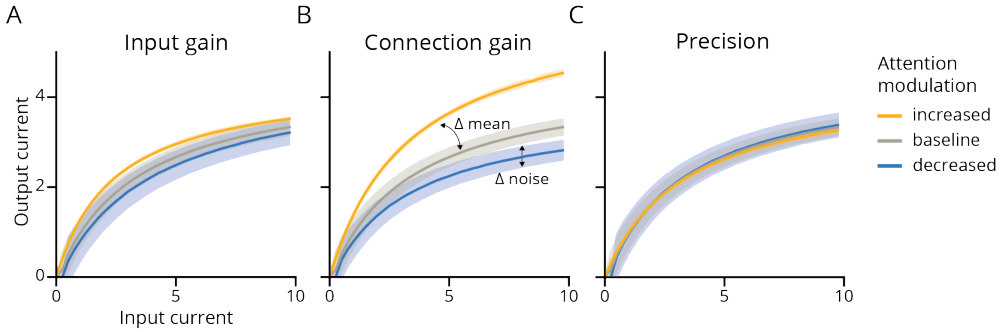

Figure 1. Overview of attention mechanisms. Attention was implemented in three different ways: as a modulation of input gain, a modulation of output or connection gain, or as precision (noise suppression). Here, the effects are illustrated based on the activation function, which represents the activity of a spiking neuron over infinite time steps. During the main experiments, also combinations of these mechanisms were evaluated. (A) For input gain, the integrated current is multiplied with a spatial weight and a gain factor, together resulting in the attention modulation. Since this manipulation happens before the non-linear process, the amplification mainly affects middle range values. (B) For connection gain, the postsynaptic weight is modulated thereby effectively changing the slope of the activation function and thus producing the largest modulations for the strongest inputs. As a by-product, this also scales the approximation noise. (C) For precision, both the adaptation speed and the post-synaptic weight were modulated. This results in comparable activation levels across the entire value range, but leads to the differences in the amount of noise.

properties after spiking conversion. In the brain, neural information transmission also incurs task-unrelated noise (e.g. Allen and Stevens, 1994), which may also affect perceptual performance (Wyart et al., 2012). One can conceive that a gain modulation might not only boost the signal during processing, but also the noise and can thus have detrimental effects depending on the signal-to-noise ratio. Therefore, having a model that also has task-unrelated noise is crucial for understanding how different attention mechanisms may affect and interact with this signal-to-noise ratio.

In the current study, we capitalized on these properties of sDCNNs to examine how attentional modulation of neural activity may enhance performance. Specifically, we directly compared effects of three main kinds of attention mechanisms on performance and network processing, namely: input gain, connection gain and noise suppression (precision). For introducing signal gain that captures the asymmetry in our activation function, we applied gain to the incoming current of a spiking unit (input gain, Fig. 1A) or we applied gain to the outgoing spike train of the spiking unit. The latter is equivalent to changing the connection strength to the postsynaptic unit (connection gain, Fig. 1B). For modelling noise suppression, we implemented a mechanism that solely interacted with the internal noise, thus selectively modulating the noise during processing without changing the gain (precision, Fig. 1C). The effects of these three attention mechanisms and their different possible combinations were evaluated in the same network during a visual search task using real-world scene images with spatial cueing. Leveraging the full observability of the networks, we also examined the effects of the different attention mechanisms on markers of attentional processing derived from well-established empirical findings in primates, including the magnitude of evoked potentials, firing rates and response latency. In a final step, we systematically disentangled the representational shifts caused by the three main mechanisms for boosting signal and reducing noise using representational similarity analysis.

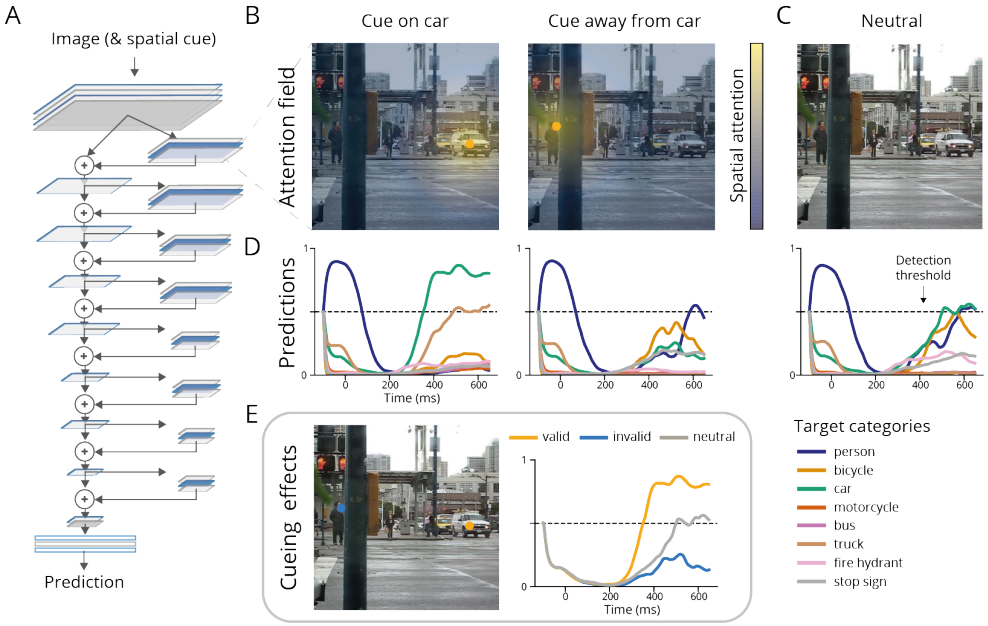

Figure 2. Using a sDCNN for naturalistic visual search with spatial cues. (A) An illustration of the ResNet18 architecture with spiking layers depicted by blue frames. The network either takes only an image or also a spatial cue as an input. We implemented the attention mechanisms only in the residual branches of the network (filled blue frames in the lower branch of the network). The output layer was a sigmoid activation function. (B) Illustration of the attention field centering on the cued location, where yellow coloured regions are being allocated more spatial attention and blue coloured less compared to baseline. (C) An example image from the curated dataset of street scenes. (D) Per image and cued location, the sDCNN produces a prediction time course indicating the presence of the eight target categories. An object is present once its prediction exceeds the detection threshold at 0.5. The x-axis shows time relative to the image onset. Passing information through the network takes approximately between 150 – 200 ms due to the spike generation. As a result, the first time points (-100 – 150 ms relative to the stimulus onset) show the biases in the network acquired during training and do not feature any information from the image yet. The left two panels show the network predictions when it was biased by a spatial cue (left: towards the location of the car, middle: away from location of a car). The right most panel show the neutral predictions for the shown image in (C). (E) Summary of the cueing effects for the car category for the predictions shown in (D). We see that with a valid cue, the network reports the presence of a car more reliably, while it misses the car with an invalid cue. The example image used in B, C and E is licensed under CC BY-NC 2.0 and was obtained from Flickr (`https://farm4.staticflickr.com/3326/3259041418_48c260317a_z.jpg`).

Results

Method's summary

We adopted a sDCNN to investigate how different attention mechanisms shape processing and performance. sDCNN are comparable to standard DCNNs in most aspects, such as their object recognition performance and training on large-scale databases. One of the most important differences is that the standard ReLU activation function is replaced during training with one that captures the activation function of a biologically-inspired neuron model. This function is not linear in the positive range and does not tend towards positive infinity, instead it saturates (see Fig. 1C for an example). A second difference is that after training with error backpropagation on ImageNet, the activation layers are replaced with more realistic spiking neurons that feature membrane potentials, adaptive thresholds and that emit temporal spike sequences as outputs. As a result, sDCNNs are evaluated over time executing the learnt weights from the training network at every time step. We here use a ResNet18-architecture (He et al., 2015) trained on ImageNet (Russakovsky et al., 2015), which was fine-tuned on a curated dataset from the COCO database (Lin et al., 2014) for testing purposes.

The sDCNNs operate on a rate code, meaning higher firing rates encode higher activation values. The binary signals passed between the layers of the network are temporal approximations of the static function learnt during training. Yet, since this is an approximation of the encoded signal, internal noise is created in every spiking layer and added to the signal. Thus the internal noise is an unavoidable consequence of using an sDCNN.

The levels of noise acquired during processing are determined by the adaptation speed of the network. We adjust the postsynaptic weight for every change in a neuron's adaptation speed to ensure that the same underlying function can still be computed, even if the adaptation speed is changed. This results in networks that still approximate the same mean function but in a more or less precise fashion depending on the chosen baseline speed of adaptation. This varying degree of noise determines the recognition performance level of the base spiking model.

We implemented seven different kinds of attention mechanisms into the same base model, thus keeping weights, architecture and internal noise levels exactly the same. For input gain, the integrated current (potential) was multiplied with a spatial attention weight and a gain factor. For connection gain, we altered the postsynaptic weight to the next layer as a function of the spatial attention weight and gain factor. For both input gain and connection gain, activations were thus increased at attended locations compared to unattended ones (Fig. 1A & B). Precision exclusively changed the noise but not the encoded values by concurrently adjusting the speed of adaptation and the spike height (Fig. 1C). We evaluated another four mechanisms, which were combinations of the first three mechanisms, for instance, a combination of input and output gain. These four combination mechanisms captured the scenario in which two or more mechanisms operate concurrently to implement attentional selection.

We evaluated all seven candidate attention mechanisms on a challenging visual search task. To this end, we curated a dataset based on the COCO database that had a homogenous context and could contain multiple target objects comparable to naturalistic visual search (person, bicycle, car, motorcycle, bus, truck, fire hydrant, stop sign, Fig. 2C). We used this dataset for fine-tuning the last fully-connected layer of the network. Contrary to training, during testing we used images that only contained one instance of the target category. The network used a sigmoid activation function in the output layer to indicate the presence of a target category (Fig. 2D).

We applied the different spatial attention mechanisms to the feature maps in the residual branches of the network (filled blue layers in the bottom row of the sDCNN,

Fig. 2A). This choice of location, and other modelling choices, were made based on separate experiments on a dataset with food scenes to avoid overfitting. The spatial reweighting was applied identically for all mechanisms. To test the effects of spatial attention, we provided the resulting networks with a valid and invalid spatial cue. A valid cue was placed on the centre of mass of the target object and an invalid cue in a distant location from the target object (Fig. 2B, E). Comparing valid and invalid cues allowed us to quantify the effect of the spatial bias introduced by the different attention mechanisms on performance and processing of the network and to compare these effects across attention mechanisms (Fig. 2E).

Prior to the main analyses, we performed a grid search to find the hyperparameters that gave the best discrimination performance when a valid cue was presented for each mechanism separately (Appendix – Hyperparameter search). This ensured that we evaluated a variant of every mechanism that was as competitive as it could be for our searched range of hyperparameters.

All attention mechanisms, except precision, produce spatial cueing in performance

Human subjects are faster and more accurate when targets occur at a validly cued spatial location compared to an invalidly cued one (Carrasco, 2011; Posner, 1980). In our first analysis, we show that all mechanisms except for precision were able to change the network's performance in the same way.

Fig. 3A shows the prediction modulation for the connection gain model (valid/invalid vs. neutral trial predictions) evolving over time. We see that the target class is modulated up or down depending on the cue's validity, suggesting that the spatial cue led to a modulation of the target class predictions and thus came at a cost or benefit in the model's prediction. A similar pattern holds for most mechanisms (Fig. S1). To quantify the effect on target discrimination of this modulation per (combination of) mechanism(s), we computed the area under the curve scores (AUC) and compared these scores to the neutral (without any spatial attention bias) and noiseless, non-spiking network (as obtained after training and before spiking conversion). For all networks, we inspected the average prediction in the period of 150 – 650 ms after stimulus onset. To assess whether the mechanisms introduced spatial cueing, we performed pairwise permutation tests for the difference between validly and invalidly cued trials. We find that all mechanisms produced different levels of target discrimination in response to the cues (all $p < .012$) with the exception of precision ($p = .152$). Models including connection gain outperformed other models. Fig. 3B shows that for the mechanisms with a significant cueing effect, valid cues produced consistently higher target discrimination than invalid cues.

Selective attention mechanisms are facing noise during neural processing. With the sDCNNs, this noise results in reduced target discrimination compared to its noiseless counterpart (Fig. 3B). To assess how the different attention mechanisms are dealing with the noise compared to the noiseless model, we tested for a difference between all mechanisms with a valid cue and the noiseless model's target discrimination. We find that only precision is significantly different from the noiseless network ($p < .001$). This thus indicates that especially gain-based mechanisms were able to overcome the noise with a valid cue.

As a next step, we evaluated if spatial cueing also influences detection time. The sDCNN were trained to report the presence of a target category by using a sigmoid function with a cutoff at 0.5 during a multi-label task. We here interpret the first time point after 150 ms at which this threshold is passed as detection time. Fig. 3C shows the detection time distribution for the different trial types for the connection gain

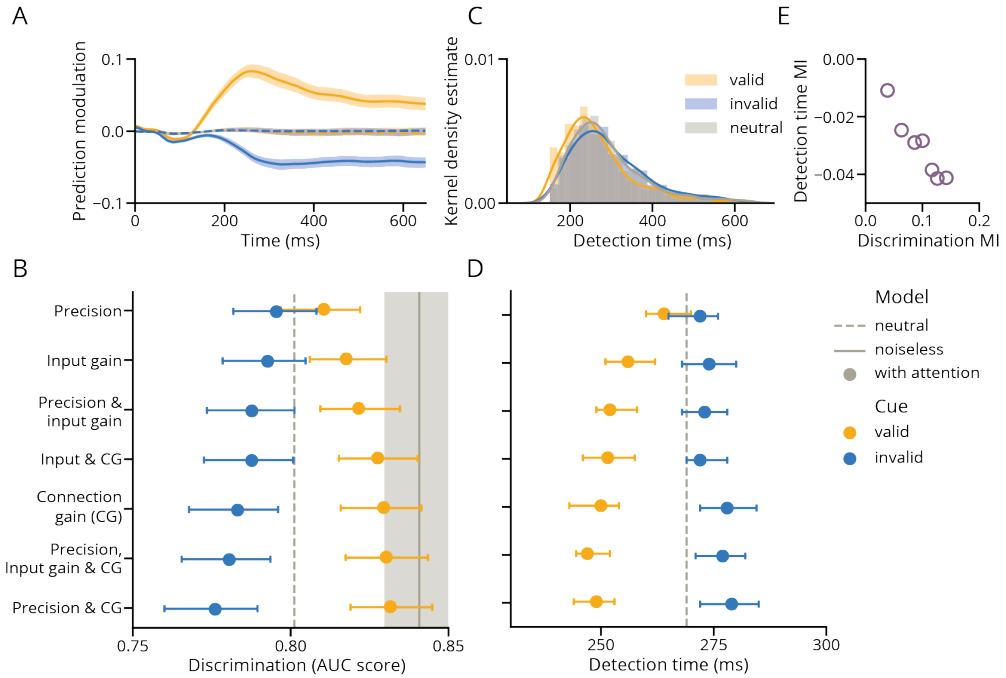

Figure 3. All (combinations of) attention mechanisms, except precision in isolation, modulate performance. (A) Example of how the network's average responses were modulated by the spatial cues by subtracting the attended trials from the neutral trials. This figure shows the data for connection gain, Fig. S1 shows the results for all other (combinations of) mechanisms. (B) Overview of spatial cueing effects on target discrimination (x-axis) for the different (combinations of) attention mechanisms (y-axis). The y-axis is ordered according to valid cue performance. The shaded area around the noiseless model as well as the error bars represent the 95% confidence interval. The attention mechanisms were defined based on a prior grid search over the gain parameter, see Appendix – Hyperparameter search. (C) Example of detection time distribution across the different trial types for connection gain. Detection time distributions for the other (combinations of) mechanisms are depicted in Fig. S2. (D) Overview of spatial cueing effects on detection times for each (combination of) attention mechanism(s). (E) Relationship between attentional modulation indices of target discrimination and detection time. More accurate models were generally associated with faster response times. The modulation index (MI) was computed as $\frac{valid-invalid}{valid+invalid}$.

mechanism (see Fig. S2 for all other mechanisms). Comparing the medians of these distributions between the valid and invalid cues reveals significant differences in detection times for most mechanisms (all $p < .001$) except for precision ($p = 0.075$, Fig. 3D). Again, models including connection gain produced the largest reaction time validity effects.

Lastly, spatial attention has been shown to similarly affect accuracy and reaction time in human observers (Carrasco and McElree, 2001; Prinzmetal et al., 2005). One may thus predict that more accurate models also generally produced faster detection times. To evaluate this, we computed the attentional modulation index ($\frac{(valid-invalid)}{(valid+invalid)}$) for target discrimination and detection times separately. While we do not want to quantify this link on the basis of seven data points, we observed a linear relationship between both performance measures (see Fig. 3E). Contrary to target discrimination, the networks were not optimized for having faster detection times when a valid cue was presented. However, all mechanisms that showed an increase in target discrimination also showed an increase in detection speed, with the better models also being the faster ones. This suggests that these processes are, just as in humans (Carrasco and McElree, 2001; Prinzmetal et al., 2005), intrinsically linked.

All attention mechanisms can qualitatively replicate neural changes

The effects of spatial attention on processing in visual cortices have been extensively documented (e.g. Maunsell, 2015). Here, we investigated how the different mechanisms implemented in our sDCNNs induced neural changes of the kind and magnitude observed in empirical studies. In particular, we examined changes in evoked potential, firing rates and latency. For this we looked at spiking units from the sixth ResNet block, which is in the middle of the network (Fig. 2A). These units still have some degree of spatial specificity while also having a sizable impact on network performance. Within the fourth block, we selected units that represented the features at the centre of mass of the target object. We recorded from these units under three conditions, without a cue, with a valid cue, and with an invalid cue.

The evoked potentials, that is, the integrated currents inside of the spiking units, are the most comparable to the activations observed in non-spiking DCNNs. If an input gain mechanism works as expected, it should modulate these values, as these are the proportional markers of encoded activation. With the sDCNN, we have the possibility to see how the encoded values at valid and invalid locations change with the different attention mechanisms, while also being able to link them directly to firing rates. In contrast, measuring the evoked potentials in single neurons simultaneously in a large population can be experimentally challenging. To obtain the evoked potential without directly measuring the attentional manipulations (e.g., a change in post-synaptic weight) as well, we integrated the spike trains into another neuron.

Fig. 4A shows an example for one mechanism for the average response across the entire dataset (connection gain, for other mechanisms see Fig. S3). Comparing the mean modulation between 150 – 650 ms across mechanisms indicates that the evoked potentials were modulated significantly by all attention mechanisms, as indicated by significant differences in potential amplitude between the valid and invalid cue conditions (all $p < .001$, Fig. 4B). Yet, as can be seen in Fig. 4B, the magnitude of this modulation varied between mechanisms and notably, in the same manner as performance. Indeed, relating the degree of modulation back to target discrimination suggests that for this layer, there appears to be a linear relationship: Networks with stronger layer modulations tend to also exhibit more strongly biased target discrimination (Fig. 4C). This is interesting because we did not select these networks

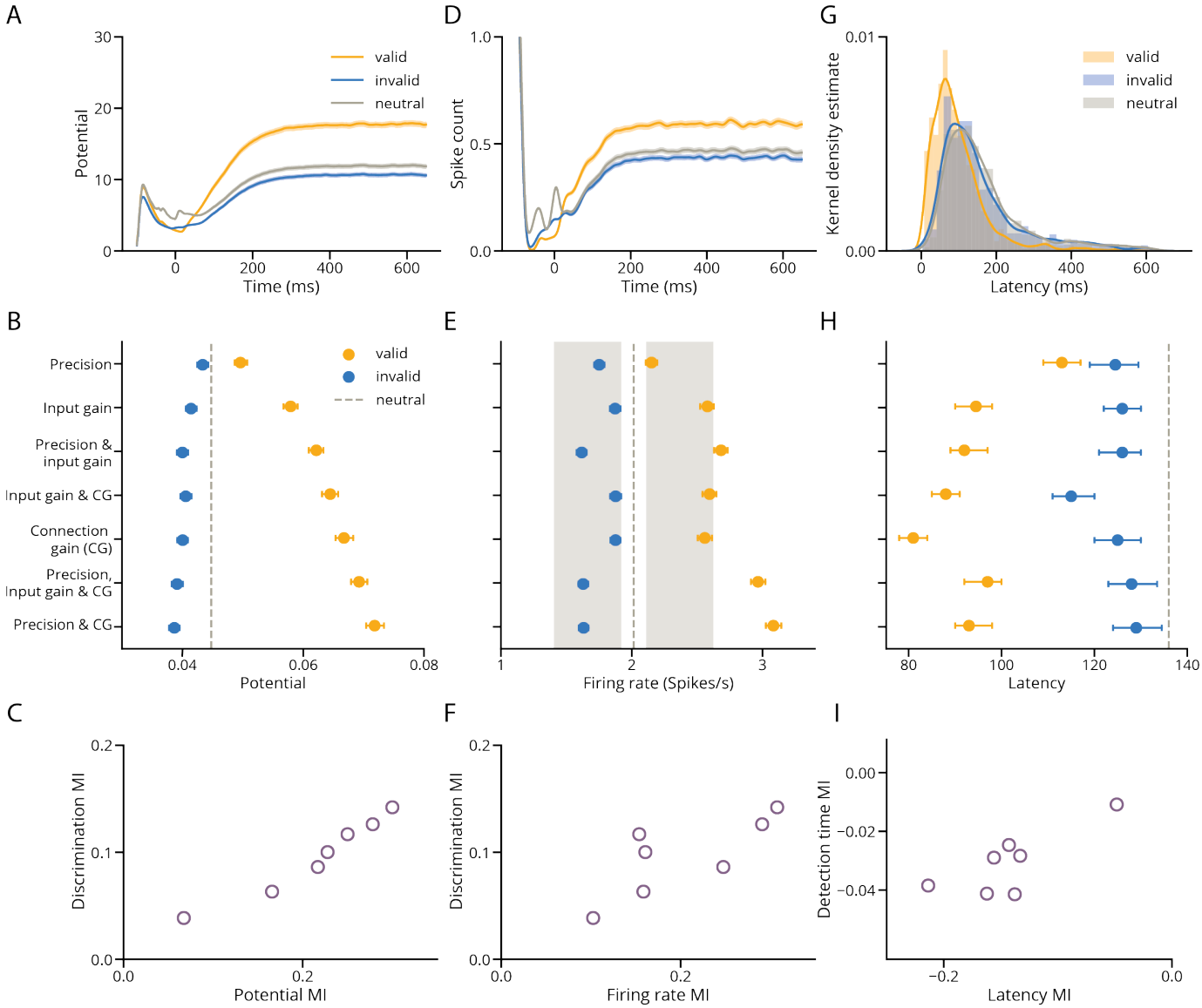

Figure 4. Effects of precision, input and output gain, and their combination on key neural indices of attention: evoked potentials, firing rate and latency. (A) Example time course of the evoked potentials measured from spiking units at the centre of the target object in a given image for the three different conditions (valid, invalid, neutral cue). The data shown is from the connection gain (CG) mechanism. The evoked potential time courses for the other mechanisms can be found in Fig. S3. Shaded areas depict the 95% confidence interval obtained with resampling (CI). (B) Mean evoked potential between 150 - 650 ms compared for the different attention mechanisms. The error bars depict the 95% CI for B, E and H. (C) Larger evoked potential modulation is generally associated with larger target discrimination modulation across the different mechanisms. (D) Example time course of the average spike count from the same neurons shown in (A). The average spike count time courses for all other mechanisms can be found in Fig. S4. The spike count time courses have been temporally smoothed with a gaussian kernel. (E) Mean firing rates for the different attention mechanisms. The grey boxes indicate a modulation of 5 - 30% from baseline as observed in experimental data (Maunsell, 2015). (F) Relationship between modulation indices of firing rates and target discrimination. (G) Example distribution of latency estimates for connection gain. Latencies were estimates from the smoothed spike count histogram and defined as the time when 50% of the first peak was reached. For the other mechanisms, please see Fig. S5. (H) Median latencies for the different conditions. (I) Relationship between the modulation index in latency and modulation of detection times across the different mechanisms.

based on similar levels of processing modulations but rather, we picked the best performing models irrespective of their parameters. In turn, this suggests that some mechanisms are better able to drive the network's activity up in service of performance.

Next, we studied how these changes in evoked potential may translate to changes in firing rate. Modulation of firing rate by spatial attention is the most common observation in electrophysiological studies. Firing rates typically change as a function of cue validity with a modulation range of $5 - 30\%$ compared to baseline (Maunsell, 2015). Here, we found that most mechanisms resulted in changes within this range, with the exception of the two best performing models.

For measuring the firing rates, we recorded spike trains at the output of the next unit. Since for some mechanisms the connection gain to the next neuron was the mechanism, this keeps the impact of attention for all measurements the same. Fig. 4D shows the average response across the entire dataset for one mechanism (connection gain, for other mechanisms see Fig. S4). In contrast to typical analysis protocols, we did not make any preselection of units based on their responsiveness. Because our networks are sparse in their activations, this results in low firing rates (Fig. 4E).

Across mechanisms, there was a significant change in firing rates in response to a valid compared to an invalid cue (all $p < .001$, Fig. 4E). To link these changes back to experimental data, we plotted the range of $5 - 30\%$ of modulation compared to the neutral model (grey shaded areas). It becomes clear that the two best performing models are producing changes that are beyond the biologically observed range. For those mechanisms that stay within this range, connection gain is the best performing mechanism. All in all, these firing rate modulations suggest that all gain mechanisms in isolation translated to comparable firing rate modulations.

Lastly, spatial attention has not only been shown to alter the response magnitude of visual cortical neurons, but also to modulate the latency of their responses. Sundberg et al. (2012) reported that attention was associated with a reduction in latency between $0.5 - 2$ ms of both the spiking and LFP responses of neurons in V4. Similar findings have also been reported for MT, suggesting this to be a phenomenon present in multiple visual areas (Galashan et al., 2013). In our next analysis, we examined the latency changes introduced by the different attention mechanisms in response to a valid and invalid cue. In brief, we find that all mechanisms markedly affected the processing latency, yet at a much larger magnitude than observed in neural data.

To obtain latency estimates, we re-analysed the firing rate data with regard to the first modulation compared to baseline activity per trial. We defined latency as the time point by which the smoothed spike density function reached 50% of the maximum firing rate of the first peak in the response after stimulus onset following Sundberg et al. (2012).

Fig. 4G shows the distribution of estimated latencies for one attention mechanism (connection gain, for other mechanisms see Fig. S5). The difference in distributions along the x-axis indicates a large decrease in latency for valid cues. Indeed, for all mechanisms, we observe faster response latencies in valid compared to invalid cue conditions (all $p < .001$, Fig. 4H). Furthermore, most mechanisms even produced a latency benefit irrespective of cue validity (all $p < 0.027$, except for precision & connection gain). This suggests that enhanced processing at some location, even if it is an invalid location, led to an overall increase in model responsiveness. We believe this could be due to the convergent architecture and large receptive fields of ResNet18.

Relating the modulation in latency in this layer back to the detection time in the networks' performance reveals a weak linear link (Fig. 4I) with the mechanisms with the fastest detection time also having a more strongly modulated latency. Yet, this association might be rather noisy because of our finding that even invalid cues reduced response latencies and given our approach of estimating latencies. In fact, our observed

latency differences are much larger than typically observed in neural processing (11.5 – 44 ms vs . 0.5 – 2 ms). This suggests that there are potentially important differences between the neural and sDCNN firing rate data in the onset response and the temporal adaptation that may have made this analysis approach less suitable for sDCNNs.

In sum, we found that all mechanisms modulated evoked potentials, firing rates and latencies according to cue validity, thereby mostly paralleling observations from neural recordings. For all three measures, we observe a qualitative link to performance, supporting that the examined layer was involved in the model's changes in performance.

Only gain mechanisms introduce additional information

Lastly, we sought to capture the impact that these various attention mechanisms have on the representations in the network. In primates, spatial and category-based attention have been reported to improve population coding in inferior temporal cortex for real-world objects among distractors (McKee et al., 2014; Zhang et al., 2011), suggesting that selective attention can amplify the signal-to-noise ratio in a noisy neuronal population. What still remains an open question is whether this improvement in population coding stems from a less noisy representation (noise suppression) or rather can be attributed to additional information about the attended object or location (signal enhancement). We here show that gain mechanisms in particular led to additional signal, while all mechanisms featured noise suppression.

With our models, we are in the position of having full access to all the units in a layer without incurring measurement noise. In addition, we also have a notion of the noise of the network at baseline (neutral network) and the representation of the noiseless network (non-spiking). With this, we can disentangle the relative contribution of noise recovery and added signal due to the spatial information of the attention cue by using representational similarity analysis (Kriegeskorte et al., 2008). For this analysis, we define noise recovery as an increase in similarity between a spiking network and its noiseless counterpart (Fig. 5B) and an increase in signal as an increase in similarity of the spiking network to a fully categorical representation (Fig. 5A).

Due to computational constraints for the sDCNN, we selected a random set of 50 images per category (400 images in total) and focused our analysis on the three key mechanisms: input gain, connection gain and precision. Since the effects on population coding have been reported in inferior temporal cortex, we chose the first spiking layer of the decoding block (Fig. 2A, the first spiking layer after the last ResNet block). We constructed a representational dissimilarity matrix (RDM) for every time step of the networks with different attention mechanisms while presented with a valid and invalid cue, respectively. Fig. 5C shows four example time frames from the neutral network to illustrate the temporal evolution of the RDMs. To quantify noise suppression and signal enhancement, respectively, we used the correlation with the noiseless (Fig. 5B) and categorical model (Fig. 5A) as our metric. To estimate whether there was a statistically significant shift in explained variance, we compared the unique correlations for a valid to the neutral and invalid condition within a given mechanism by constructing a 95% confidence interval for the difference between two conditions and estimating variability by means of bootstrapping with replacement.

Examining the changes in RDMs over time across the different mechanisms and cues, we see an increasing correlation to the noiseless network over time that stabilizes around 300 ms after stimulus onset (Fig. 5D, left panel). Comparing the three different attention mechanisms to the neutral network after controlling for categorical information (Fig. 5D, middle panel), we observe that all mechanisms became more similar to the noiseless network upon presentation with a valid cue (all $p < .002$). When examining the effect of an invalid cue (Fig. 5D, right panel), only input gain and connection gain led to a decrease in similarity (all $p < .002$, $p_{precision} = .02$). Pairwise

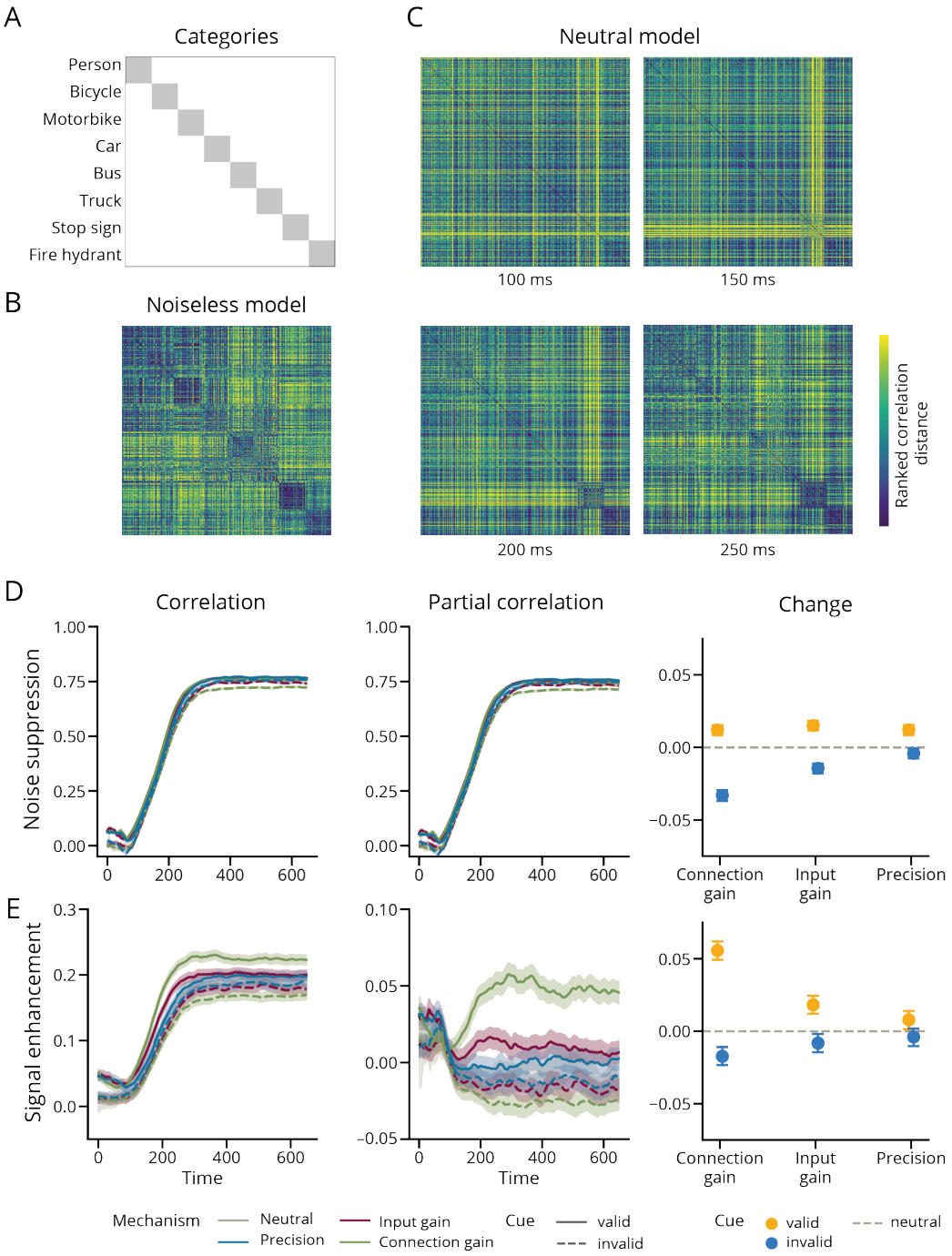

Figure 5. Gain mechanisms selectively enhance the categorical information. (A) Organization of the categories in the representational dissimilarity matrix (RDM). For every category, there were 50 images. The categories were organized according to visual similarity. For illustration purposes, the correlation coefficients were ranked. (B) RDM of the layer of interest from the noiseless, non-spiking network. (C) RDMs from four timepoints after stimulus onset for the neutral network. (D) Left panel: Pearson-correlation coefficients between the noiseless model and the temporal RDMs separately for the three attention models (precision, input gain, connection gain) and the neutral model. The shaded areas represent the 95% confidence interval obtained from resampling. Middle panel: Partial correlation coefficients for the comparison in the left panel and now controlling for the categorical RDM shown in (A). Right panel: Average difference in partial correlation between attention models and the neutral model. (E) Left panel: Pearson-correlation between the categorical model and the temporal RDMs shown separately for the three attention models and the neutral model. Middle and right panel: Same as in D, but now controlling for the noiseless model.

comparisons within a cue condition indicate that all mechanisms were different from one another (all $p <.002$) except for precision and connection gain with a valid cue ($p = .836$). Altogether, this suggests a reinstatement of the noiseless network representations for all mechanisms if presented with a valid cue and a decrease in similarity for invalid cues for gain mechanisms.

For signal enhancement as expressed in a correlation with the categorical model, we observed similar temporal characteristics as for the noise recovery (Fig. 5E, left panel). Yet, after controlling for the effect of noise recovery, we found that only the mechanisms that changed the signal featured significantly more additional categorical information with a valid cue (Fig. 5E, middle & right panel, $p_{precision} = .012$, remaining $p < .002$). The same was true for an invalid cue, where only input gain and connection gain decreased the amount of categorical information compared to the neutral network ($p_{precision} = .012$, remaining $p < .002$). Pairwise comparisons between the mechanisms within a validity condition further support this by indicating differences between all mechanisms for both valid and invalid cues (all $p < .002$).

In sum, the different attention mechanisms introduced different representational changes: While all examined mechanisms reinstated the representations used by the noiseless network to a similar degree, mechanisms that targeted the signal were more effective at adding categorical information beyond the trained weights compared to the noiseless network. This suggests that additional categorical information was derived from the spatial cue that effectively helped performance.

Discussion

In this modelling study, we examined how spatial attention may affect sensory processing and performance on a challenging visual search task using a sDCNN. Specifically, we directly contrasted effects of three different main mechanisms previously proposed to subserve selective attention (e.g. Dayan et al., 2000; Feldman and Friston, 2010; Martinez-Trujillo and Treue, 2004; Reynolds and Heeger, 2009), namely gain modulation on the input to a neuron, gain modulation on its post-synaptic connection or noise suppression in the form of precision. We found that connection gain was the most effective at implementing spatial attention, as indicated by the largest performance modulations, whereas precision was not effective. That is, our gain modulations were associated with both faster and more accurate performance when targets occurred at a validly cued spatial location compared to an invalidly cued one, just as humans are (Carrasco, 2011; Posner, 1980). Moreover, connection gain best reproduced several key experimental findings in visual cortex (Maunsell, 2015), including a proportionate modulation in firing rates. Disentangling the representational changes introduced by the three main mechanisms in the network revealed that gain-based mechanisms in particular added task-relevant information, while all networks showed similar recovery from the noise. These latter results mirror findings from the animal and human literature that show that attention can enhance the representational content of neural activity (e.g. Jehee et al., 2011; Zhang et al., 2011). Together, these findings advance our understanding of how spatial attention might be mechanistically implemented at the neural level, as discussed in more detail below.

Our finding that output gain was not only more effective than precision, but also than input gain, highlights that the asymmetry in the activation function plays a big role for the efficacy of gain on the model's performance: output gain was more effective at enhancing activation in a useful fashion and thereby biased the network's performance more strongly compared to input gain. This result helps us to better understand how this asymmetry can act as a constraint for a gain mechanism. This finding also sheds new light on past studies that used ReLUs to implement gain

(Lindsay and Miller, 2018; Luo et al., 2020): A shared feature between their gain implementation and our connection gain is that both approaches specifically boost larger values, that is, higher values are proportionally more affected compared to smaller values. Based on our results, we can thus speculate that these larger values in particular are important for boosting performance, at least in DCNNs.

In contrast to gain, we found that precision was not effective at implementing attentional selection as reflected in performance measures. This finding was unexpected since global changes in precision have been shown to lead to substantial changes in noise during processing in a previous sDCNN study and are a powerful way to increase and decrease the network's performance (Zambrano et al., 2018). It also does not align with notions that assign an important role to reliability or precision in selective information processing, and that emphasize noise reduction rather than signal amplification (Dayan et al., 2000; Feldman and Friston, 2010; Parr and Friston, 2017). One potential explanation that could reconcile our finding with this past work is that we did not use precision in the context of a Bayesian observer framework (Feldman and Friston, 2010; Parr and Friston, 2017). Instead, our notion of precision is directly grounded in the signal-detection framework, where a signal is represented in a more or less narrow distribution, making it thereby more or less distinguishable from noise. It is a possibility that such a mechanism proves to be more effective when used on prediction errors rather than on stimulus-driven information. Noise reduction may also play a more dominant role in well trained systems: a set of event-related potential (ERP) studies by Itthipuripat et al. (2014, 2017) showed that spatial attention was associated with a gain modulation of early visual-evoked potentials early in training, but with noise reduction at advanced stages of training (Itthipuripat et al., 2017). These findings suggest that the mechanisms through which spatial attention facilitates performance depend on the specific behavioural training regime used, with gain-type mechanisms subserving selective information processing in relatively untrained systems, as in our models. Under the assumption that learning affects the synaptic strength akin to a momentary gain change, we can make the prediction that precision should be especially successful in our models once we allow for weight adaptation after trials with attentional selection. Future studies are necessary to test this prediction.

Our findings on input and connection gain raise the question of how our examined attention mechanisms might link to selective attention in biological neurons. The effects of input gain could be related to what has been referred to as contrast gain in neural processing (Reynolds et al., 2000) and such neural changes are mathematically equivalent to a multiplication of the incoming current (Reynolds and Heeger, 2009). Yet, how specifically this multiplicative gain could be implemented in a neuron is still matter of active research (for a recent review, see Ferguson and Cardin, 2020). The effects of connection gain resemble neural changes described as response gain: increased firing rates that scale beyond the maximum response observed under neutral conditions (McAdams and Maunsell, 1999; Treue and Martínez Trujillo, 1999). This gain profile has been argued to arise from the same neural populations as contrast gain, yet in situations in which a relatively small attention field is paired with a large stimulus (Reynolds and Heeger, 2009). For its biological implementation, it has been suggested that both neural contrast and response gain can be obtained by combining a multiplicative gain on excitatory inputs with lateral or feedforward inhibition (Beuth and Hamker, 2015). Another recent proposal is the addition of a current in a recurrently connected excitatory-inhibitory circuit, which results in a multiplicative gain that can also show a switch from contrast to response gain (Lindsay et al., 2019). An alternative for how connection gain effects could be implemented independently of contrast gain is as a change in synaptic efficacy, as was reported in a set of studies in LGN and V1 (Briggs et al., 2013; Hembrook-Short et al., 2018). Future experimental as

well as modelling studies will be crucial to further link biological circuits to computational functions during attentional selection. Our results suggest that there is a computational advantage in amplifying a neuron's outputs, akin to both response and connection gain, rather than its inputs, even when the system features noise.

A central goal of the current modelling experiments was to study attention mechanisms in the context of more biologically plausible processing constraints. A challenge inherent to any such endeavour is that the validity of findings is determined by the accuracy of the model in capturing relevant biological properties. Our study was conducted with adaptive spiking neurons (Bellec et al., 2018; Bohte, 2012). We chose this neuron model because it allowed us to implement key biological properties (e.g. adaptive thresholds, saturation at realistic firing rates, Zambrano et al., 2018) and relate changes in firing rates, population information, neural latencies, discrimination performance and detection times in the network's output to key findings from behavioural and neuroscientific studies of attention. While thereby providing important additional, neurally grounded constraints for evaluating the feasibility of different mechanisms compared to existing DCNNs that have used ReLUs, our neuron models do not have stochastic firing thresholds or other sources of inter-trial variability due to the intended conversion after training. As a consequence, it is not possible to relate our results to effects of selective attention on spike count variability or noise, that have also been reported in the literature as a measure of noise (Cohen and Maunsell, 2009; Mitchell et al., 2007, 2009). A recent study suggests that such measures can be reproduced in a circuit of neuron models injected with synaptic noise (Lindsay et al., 2019). Future studies that include such synaptic noise are necessary to also establish a more direct link to noise correlations and to complement our here presented results.

In this study, we examined how three main attention mechanisms can shape a complex and noisy process such as object recognition in natural scenes. Leveraging sDCNNs, we were able to inspect the computational consequences of different proposals of how selective attention may be implemented in the brain. Across a variety of measures, we observed a computational advantage of gain-based and in particular connection gain-based mechanisms, in contrast to noise reduction. Our results highlight that sDCNNs provide a suitable modelling framework for connecting empirical observations from performance to neural processing and illustrate how they can be used to differentiate between theories in the context of richer tasks.

Methods & Materials

We implemented different spatial attention mechanisms (see Attention mechanisms) into a spiking deep convolutional neural network (see Spiking deep convolutional neural networks (sDCNNs)) to understand the effects of these mechanisms on performance and processing (see Analyses) of these networks during a visual search task with spatial cues (see Visual search task & dataset).

Visual search task & dataset

For the visual search task, we used street scenes as stimuli that could contain eight possible target categories. Furthermore, we used a dataset with food scenes during model development and for exploratory experiments. To obtain images with a set of potential target categories sharing a context, we first curated these two datasets from the Common Objects in Context database (COCO, Lin et al., 2014). To obtain a homogeneous context, we quantified similarity in context as similarity in the stuff-annotations of the COCO dataset (Caesar et al., 2016). Specifically, we focused on the super categories from the stuff-annotations (total: 15, e.g. wall, ceiling, sky, water)

and defined a vector specifying the presence for all these stuff-annotations for every image. From this we computed the Euclidean distance between different target category centres (average of all individual vectors belonging to the same target category): If a target category on average has the same stuff as another target category, these would have a small distance in such a stuff space. Based on this procedure, we identified a cluster with low distances in stuff-space for street scenes that contained 12 categories (person, bicycle, car, motorcycle, bus, truck, traffic light, fire hydrant, stop sign, parking meter, bench, dog) and for food scenes containing 16 categories (bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, pizza, donut, cake).

Next, these categories were further processed to select high-quality images with only one recognizable instance of a target category present, to make the spatial cue informative and advantageous for performance. For the street scenes, we selected images that had a street stuff annotation, for the food scenes, there could be a table, cloth, food-other, vegetable, salad, fruit, and/or napkin stuff annotation. For both datasets, we selected images with target objects that were big enough (> 0.05 % of the image), placed in a not too complex scene (spatial coherence < 1.2 based on Scholte et al., 2009), were not too central (outside of a radius of 5% from the image center) and salient enough (summed object probability density from DeepGaze II > 0.04, Kümmerer et al., 2016). This resulted in 8 eligible target categories with at least approximately 50 images with a single target object (street scenes: person, bicycle, car, motorcycle, bus, truck, traffic light, fire hydrant, stop sign, parking meter; food scenes: bottle, wine glass, cup, cutlery, bowl, sandwich, carrot, cake). To obtain a sufficient number of images (around 50 images) for 8 categories in the food dataset, we combined the categories knife, spoon and fork into a single category of cutlery. The code for recreating these datasets is available at `https://github.com/lynnsoerensen/SpatialAttention_sDCNN_2020`. Due to the datasets' license status, the final datasets will not be shared publicly but can be requested directly from the authors.

The visual search task was performed on the thus obtained single-target images from the street dataset (N=1628, 224x224 pixels). For every image, we defined a valid spatial cue based on the centre of mass of the target object and an invalid cue pointing to an irrelevant location. The invalid locations were obtained by randomly sampling from a uniform distribution that was constrained to the minimum and maximum values observed for the valid cues (0.1 - 0.94 for the horizontal and 0.03 - 0.95 for the vertical extent of the image). This sampling process was repeated until the invalid location was at least 0.5 of the image extent away from the valid cue. The valid and invalid locations had on average a distance of 0.62 of the image extent. Due to imperfect COCO annotations, some images also featured two instances of the same target category. To quantify the effect of the spatial bias introduced by the attention mechanisms, we compared performance between a validly cued (centre of mass of a target object in the scene), an invalidly cued (an unrelated location) and an uncued (neutral) processed dataset during most analyses.

Spiking deep convolutional neural networks (sDCNNs)

Training & Fine-tuning

We used a first generation ResNet18-architecture (He et al., 2015) in which the rectified linear units were replaced with a custom activation function approximating a spiking neuron's in- and output relationship, which takes the form of a rectified sigmoid-like function. The network implementation is based on (Zambrano et al., 2018) and details and derivations of the activation function can be found there. This architecture was trained on the ImageNet dataset (Russakovsky et al., 2015) with stochastic gradient

descent (initial learning rate: 0.1 with Nesterov momentum of 0.9, decay: 0.0001). The training parameter choices closely followed He et al. (2015) (training augmentation: random cropping and horizontal flipping, test augmentation: center crop). The learning rate was divided by 0.1 every 30 epochs. The final model performed at 64.04% (Top-1 accuracy) on the ImageNet validation set.

The pre-trained network was fine-tuned on the street dataset by replacing the last fully connected layer by one with 8 units and a sigmoid activation function. All remaining layers and their weights were kept unchanged. For this step, we used images with more than one target object present selected based on less stringent criteria as the test images (only based on stuff-annontations) resulting in 8640 training and 2160 validation images. The newly added weights were optimized with a binary cross-entropy loss, a learning rate of 0.0001 and an ADAM optimizer for 100 epochs. The final binary accuracy on the multi-object dataset was 88.74%.

Spiking inference

After training and fine-tuning, the trained weights were evaluated with a spiking network that computes continuously over time, encoding its activations in binary signals. Practically, this means that the formerly learned activation function approximation is replaced with a layer of spiking neuron models. These spiking neuron models feature firing thresholds, membrane potentials and refractory responses, and they integrate and decay current over time (Bohte, 2012). This allowed us to convert a binary signal over time into a continuous one, making them effectively an analog-to-digital converter (Lazar and Toth, 2003; Yoon, 2017). Based on this approach, it is possible to maintain a network performance that is close to the state-of-the-art (Rueckauer et al., 2016; Zambrano et al., 2018), while using biologically plausible firing rates. All DCNNs, as well as sDCNNs models were implemented, trained and evaluated in Keras with a tensorflow backend. The code is available at `https://github.com/lynnsoerensen/SpatialAttention_sDCNN_2020`.

For all experiments, we chose a baseline firing threshold of 0.45 resulting in a network in which every neuron fires on average at 18.63 Hz. We had a 100 ms pre-stimulus period to dissociate the adaptation in the network, such as the saturation of the adaptive threshold, from the image onset response. The network was evaluated over 750 ms.

Attention mechanisms

We tested different attention mechanisms to see if they can implement spatial attention in the sDCNN. The three families of mechanisms we examined here were precision, input gain and connection gain. Unless stated otherwise, all aspects were kept identical across mechanisms.

In line with earlier work (Anton-Erxleben and Carrasco, 2013; Reynolds and Heeger, 2009), we modelled the distribution of spatial attention R as a bivariate gaussian distribution over space (Fig. 2D). The centre of the gaussian was placed at the cued location. The standard deviations were kept at 40 pixels for both spatial dimensions. We chose this standard deviation because the average area of an object is equivalent to a circle with ca. 41 pixel radius. We normalized the attention field to have an average of 0 over all spatial locations based on the assumption that attention involves a redistribution of resources.

All mechanisms targeted the spiking layers in the residual branch of the network (Fig. 2B) and the skip branch was thus unaffected by the attentional modulation. This choice was made based on exploratory experiments on the food dataset (see Visual search task & dataset), in which we observed that these branches are set up to be antagonistic, thus

modulating the skip branch while also targeting the residual branch cancelled out the effects of attention. In these exploratory experiments, we also observed that it was the most effective to target all residual branches simultaneously. To do this, the attention field was downsampled to match the spatial dimension of the residual branches.

The implementation of input gain to the spiking layer followed

$$S(t) = (\phi * I * (R * \alpha_{input} + 1))(t)$$

where $S(t)$ is the integrated current at time t, I is the incoming current, ϕ corresponds to the membrane filter, R is the attention field, α_{input} is the gain factor. This thus scales the incoming activation of the neuron and also adjusts the spike production accordingly.

For connection gain, a gain factor was applied to the outgoing synaptic connection, which can also be expressed as a change in spike height:

$$spike(t) = spike(t) * h * (R * \alpha_{output} + 1)$$

where $spike(t)$ codes whether there was a spike event or not at time t, h is the spike height that is determined by the adaptation speed mf and α_{output} is the gain factor. This step scales the out-going spike trains with regard to the impact they will have on the next layer.

For precision, the speed of the adaptation mf was modulated by the spatial attention field:

$$mf = mf_{baseline} * (1 - R)$$

This defines speed of adaptation for a spatial location that is modulated by the spatial reweighting of R. When also adjusting the spike height h, this produces a situation, in which the mean approximation of the neuron stays the same, yet the precision of the approximation is varied (Fig. 1C).

To compare the efficacy of these different mechanisms, we wanted to obtain well-performing models for all mechanisms. This entailed computing various versions while varying α_{input} and α_{output} as free parameters and either adding precision modulation or not. For the main experiments, we chose the best performing model with a valid cue. We searched parameters between $0 - 0.5$ in increments of 0.05 for the isolated parameter and $0 - 0.3$ in increments of 0.05 for the combinations. The final hyperparameter used for connection gain was 0.3, for input gain it was 0.15, for their combination it was 0.2 for connection gain and 0.05 for input gain. For all mechanisms, these parameters were the same for a network with and without added precision.

Analyses

Model performance - target discrimination & detection times

The analyses on model performance were done on the output of the sigmoid activation function obtained from the sDCNN. We evaluated the model's target discrimination with the area under curve (AUC) metric across all experiments. To convert the spiking model predictions to this metric, we computed the average of the prediction time-course between 150 and 650 ms after stimulus onset.

The attention conditions (valid, invalid, neutral) were statistically compared with permutation-testing for a difference in average AUC. This was done separately for every mechanism. We performed four pairwise comparisons per mechanism and adjusted our alpha levels according to the Benjamini-Hochberg procedure per inference (Benjamini and Hochberg, 1995) for this analysis and for all later analyses.

The model was optimized to predict the presence of a target class by returning values larger than 0.5 due to the use of a sigmoid function at the output layer. We used

this feature to estimate the detection time of the model. We defined the detection time at the time point at which the target prediction time course crosses the detection threshold. Prediction time courses were analysed 150 ms after stimulus onset to separate the stimulus response from the general adaptation response caused by the biases in the network (see view2C for an example).

For both AUC scores and detection times, we estimated 95% confidence intervals using stratified sampling with replacements across trials.

Prediction modulation was estimated by subtracting the neutral models' trials responses from the valid or invalid trial responses, thereby showing the actual change in prediction introduced by the spatial cue. The 95% confidence intervals were obtained by sampling with replacement from the trials per time point.

Layer responses – evoked potentials, firing rates & latencies

To obtain a better understanding of the changes introduced by the different mechanisms, we analysed a layer's response more carefully. Specifically, we analysed the spiking layer of the sixth residual branch (layer 25, 14x14x256). All analyses attempted to follow approaches from neural recordings as closely as possible.

For all measures, we recorded units representing the centre of mass of the target object for all feature maps in a layer. We chose to do this, because we did not want to make a sub-selection of units based on a small sample of stimuli but instead our goal was to look at the entire layer to get a representative sample. Attention conditions were compared by recording from the same set of units while either altering the attention field according to the spatial cue or by having no attention modulation at all for neutral.

As outlined above, some attention mechanisms targeted the outgoing synaptic connections, thus measuring the targeted units directly would have the side effect of measuring this very manipulation. We therefore added the spikes of the manipulated unit into another unweighted unit.

For the evoked potentials, we integrated the spikes from the manipulated neurons and read out the potential. These are thus a direct precursor of the firing rates. For the firing rates, we reported the output of this unweighted neuron. For illustration purposes, the obtained spike histograms across all features maps were smoothed with a temporal gaussian kernel (standard deviation of 8ms) in Fig. 4D.

The response latency for the spike responses were calculated by a metric that closely followed Sundberg et al. (2012) and Lee et al. (2007). Specifically, latency was defined as the time by which the spike density function reached 50% of the maximum firing rate of the first peak in the transient response after stimulus onset.

To estimate the latency, the spike responses were first smoothed over time with a gaussian filter of 8 ms. In a next step, every trial was compared back to baseline measurements in which 16 average images based on 50 randomly picked images from the training set were presented to the network. Based on the activation of the baseline activation in every trial, we determined the 99.99% percentile of activation (corresponding to 3.72 SEM in Sundberg et al., 2012) and used this value as a criterion to identify the first local peak after stimulus onset. Both baseline and experimental trials were baseline-corrected based on the 50 ms prior to image onset.

If a trial did not surpass the criterion for activation, we could not estimate the latency, which led to a different amount of trials that were excluded per attention mechanism and cue (see Table 1) of a total of 1628 trials for every cue condition. For all measures, we estimated 95% confidence intervals by resampling with replacements across trials.

Table 1. Missing trials for latency estimation per attention mechanism and cue. The procedure of the latency estimation involved to assess whether a spiking unit was more active compared to baseline. If this baseline criterion was not exceeded, no latency estimate was made. More active networks are thus less likely to have missing trials compared to inactive networks.

Mechanism	**Valid cue**	**Invalid cue**
Precision	127	203
Input gain	73	191
Precision & input gain	71	234
Input & connection gain	48	129
Connection gain	43	196
Precision, input & connection gain	66	259
Precision & connection gain	62	325
Baseline (no cue)	231	

Representational similarity analysis

Due to computational constraints, we restricted the representational similarity analysis to the three main mechanisms (connection gain, input gain and precision) and we used a subset of the images compared to the rest of the experiments (400 in total, 50 per category). The resulting sub-dataset was classified by the noiseless model with a similar level of performance as other randomly drawn samples from the dataset (AUC: 0.8452 vs. 0.8409, selected vs. resampled).

For every image, mechanism and attention condition, we obtained the evoked activations for network layer 37. We chose this layer because it is the very last spiking layer of the ResNet blocks and is not targeted by attention itself since it is part of the skip branch.

We obtained the pairwise Pearson correlation between every image for every time point and model with the respective mechanism and attention condition separately, thus giving us seven representational dissimilarity time courses.

The goal for this analysis was to understand how the representations in the network are altered by the different attention mechanisms. The reasoning was that a closer similarity to the non-spiking, noiseless network means that the spiking network recovers from the introduced noise due to the spatial attention cue. In a similar vein, we expected that the similarity with the categorical RDM should grow if there is additional target object information present that helps the network at the task. To test this, we correlated every timepoint of the seven RDM time courses with either the noiseless or categorical RDMs. Due to a high correlation between these factors, we then calculated the partial correlation for these two predictors, thus giving us the unique contribution of both predictor RDMs, while keeping the influence of the other predictor constant. To compare this to the neutral model, we subtracted the partial correlations of neutral condition from all other partial correlations. We assessed statistical significance by comparing the mean bootstrapped difference between 150 – 650 ms to zero. For all time courses, we estimated 95% confidence intervals by resampling with replacements across the RDMS per timepoint.

Python packages

In addition to custom code (`https://github.com/lynnsoerensen/SpatialAttention_sDCNN_2020`), the results presented here were obtained while relying on the following Python packages: NumPy

(Harris et al., 2020), keras (Chollet, 2015), TensorFlow (Abadi et al., 2016), Pandas (McKinney and Others, 2010), Pingouin (Vallat, 2018), Scikit-Learn (Pedregosa, 2011) and SciPy (Virtanen et al., 2020). Data visualization was done using matplotlib (Hunter, 2007) and, in particular, seaborn (Waskom et al., 2020).

Acknowledgements

The authors would like to thank Leon Reteig for his advice on data visualization.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-Scale machine learning on heterogeneous distributed systems. Mar. 2016.

C. Allen and C. F. Stevens. An evaluation of causes for unreliability of synaptic transmission. *Proc. Natl. Acad. Sci. U. S. A.*, 91(22):10380–10383, Oct. 1994.

K. Anton-Erxleben and M. Carrasco. Attentional enhancement of spatial resolution: linking behavioural and neurophysiological evidence. *Nat. Rev. Neurosci.*, 14(3): 188–200, Mar. 2013.

A. Ayaz and F. S. Chance. Gain modulation of neuronal responses by subtractive and divisive mechanisms of inhibition. *J. Neurophysiol.*, 101(2):958–968, Feb. 2009.

G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass. Long short-term memory and learning-to-learn in networks of spiking neurons. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 787–797. Curran Associates, Inc., 2018.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 1995.

F. Beuth and F. H. Hamker. A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. *Vision Res.*, 116(Pt B):241–257, Nov. 2015.

S. Bohte. Efficient Spike-Coding with multiplicative adaptation in a spike response model. *Advances in Neural Information Processing Systems 25*, pages 1844–1852, 2012.

F. Briggs, G. R. Mangun, and W. M. Usrey. Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. *Nature*, 499(7459):476–480, July 2013.

H. Caesar, J. Uijlings, and V. Ferrari. COCO-Stuff: Thing and stuff classes in context. Dec. 2016.

M. Carrasco. Visual attention: the past 25 years. *Vision Res.*, 51(13):1484–1525, July 2011.

M. Carrasco and B. McElree. Covert attention accelerates the rate of visual information processing. *Proc. Natl. Acad. Sci. U. S. A.*, 98(9):5363–5367, Apr. 2001.

F. Chollet. keras. 2015.

M. R. Cohen and J. H. R. Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.*, 12(12):1594–1600, Dec. 2009.

P. Dayan and L. F. Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. pure.mpg.de, 2001.

P. Dayan, S. Kakade, and P. R. Montague. Learning and selective attention. *Nat. Neurosci.*, 3 Suppl:1218–1223, Nov. 2000.

M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, 152:184–194, May 2017.

H. Feldman and K. J. Friston. Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.*, 4:215, Dec. 2010.

K. A. Ferguson and J. A. Cardin. Mechanisms underlying gain modulation in the cortex. *Nat. Rev. Neurosci.*, Jan. 2020.

F. O. Galashan, H. C. Saßen, A. K. Kreiter, and D. Wegener. Monkey area MT latencies to speed changes depend on attention and correlate with behavioral reaction times. *Neuron*, 78(4):740–750, May 2013.

U. Güçlü and M. A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.*, 35(27): 10005–10014, July 2015.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825): 357–362, Sept. 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. Dec. 2015.

J. R. Hembrook-Short, V. L. Mock, W. M. Usrey, and F. Briggs. Attention enhances the efficacy of communication in V1 local circuits. *J. Neurosci.*, Dec. 2018.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science Engineering*, 9(3):90–95, May 2007.

S. Itthipuripat, E. F. Ester, S. Deering, and J. T. Serences. Sensory gain outperforms efficient readout mechanisms in predicting attention-related improvements in behavior. *J. Neurosci.*, 34(40):13384–13398, Oct. 2014.

S. Itthipuripat, K. Cha, A. Byers, and J. T. Serences. Two different mechanisms support selective attention at different phases of training. *PLoS Biol.*, 15(6):e2001724, June 2017.

J. F. M. Jehee, D. K. Brady, and F. Tong. Attention improves encoding of task-relevant features in the human visual cortex. *J. Neurosci.*, 31(22):8210–8219, June 2011.

K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.*, 2019.

S. M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.*, 10(11):e1003915, Nov. 2014.

T. C. Kietzmann, P. McClure, and N. Kriegeskorte. Deep neural networks in computational neuroscience. In *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press, Jan. 2019a.

T. C. Kietzmann, C. J. Spoerer, L. K. A. Sörensen, R. M. Cichy, O. Hauk, and N. Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U. S. A.*, 116(43):21854–21863, Oct. 2019b.

N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, 2:4, 2008.

J. Kubilius, S. Bracci, and H. P. Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.*, 12(4):e1004896, 2016.

J. Kubilius, M. Schrimpf, A. Nayebi, D. Bear, D. L. K. Yamins, and J. J. DiCarlo. CORnet: Modeling the neural mechanisms of core object recognition. Sept. 2018.

M. Kümmerer, T. S. A. Wallis, and M. Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. Oct. 2016.

A. A. Lazar and L. T. Toth. Time encoding and perfect recovery of bandlimited signals. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*., volume 6, pages VI–709, Apr. 2003.

J. Lee, T. Williford, and J. H. R. Maunsell. Spatial attention and the latency of neuronal responses in macaque area V4. *J. Neurosci.*, 27(36):9632–9637, Sept. 2007.

T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Lawrence Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. May 2014.

G. W. Lindsay. Attention in psychology, neuroscience, and machine learning. *Front. Comput. Neurosci.*, 14:29, 2020.

G. W. Lindsay and K. D. Miller. How biological attention mechanisms improve task performance in a large-scale visual system model. *Elife*, 7, Oct. 2018.

G. W. Lindsay, D. B. Rubin, and K. D. Miller. A simple circuit model of visual cortex explains neural and behavioral aspects of attention. 2019.

X. Luo, B. D. Roads, and B. C. Love. The costs and benefits of Goal-Directed attention in deep convolutional neural networks. Feb. 2020.

W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nat. Neurosci.*, 9(11):1432–1438, Nov. 2006.

J. Martínez-Trujillo and S. Treue. Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron*, 35(2):365–370, July 2002.

J. C. Martinez-Trujillo and S. Treue. Feature-based attention increases the selectivity of population responses in primate visual cortex. *Curr. Biol.*, 14(9):744–751, May 2004.

J. H. R. Maunsell. Neuronal mechanisms of visual attention. *Annu Rev Vis Sci*, 1: 373–391, Nov. 2015.

C. J. McAdams and J. H. Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.*, 19(1):431–441, Jan. 1999.

J. L. McKee, M. Riesenhuber, E. K. Miller, and D. J. Freedman. Task dependence of visual and category representations in prefrontal and inferior temporal cortices. *J. Neurosci.*, 34(48):16065–16075, Nov. 2014.

W. McKinney and Others. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. conference.scipy.org, 2010.

J. F. Mitchell, K. A. Sundberg, and J. H. R. Reynolds. Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron*, 55(1): 131–141, July 2007.

J. F. Mitchell, K. A. Sundberg, and J. H. R. Reynolds. Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron*, 63(6):879–888, Sept. 2009.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. Jan. 2010.

K.-I. Naka and W. A. H. Rushton. S-potentials from luminosity units in the retina of fish (cyprinidae). *J. Physiol.*, 185(3):587–599, 1966.

T. Parr and K. J. Friston. Working memory, attention, and salience in active inference. *Sci. Rep.*, 7(1):14678, Nov. 2017.

F. Pedregosa. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12: 2825–2830, 2011.

M. I. Posner. Orienting of attention. *Q. J. Exp. Psychol.*, 32(1):3–25, Feb. 1980.

W. Prinzmetal, C. McCool, and S. Park. Attention: reaction time and accuracy reveal different mechanisms. *J. Exp. Psychol. Gen.*, 134(1):73–92, Feb. 2005.

J. H. R. Reynolds and D. J. Heeger. The normalization model of attention. *Neuron*, 61 (2):168–185, Jan. 2009.

J. H. R. Reynolds, T. Pasternak, and R. Desimone. Attention increases sensitivity of V4 neurons. *Neuron*, 26(3):703–714, June 2000.

B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, and K. P. Kording. A deep learning framework for neuroscience. *Nat. Neurosci.*, 22(11): 1761–1770, Nov. 2019.

A. L. Rothenstein and J. K. Tsotsos. Attentional modulation and selection–an integrated approach. *PLoS One*, 9(6):e99681, June 2014.

B. Rueckauer, I.-A. Lungu, Y. Hu, and M. Pfeiffer. Theory and tools for the conversion of analog to spiking convolutional neural networks. Dec. 2016.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, Dec. 2015.

H. S. Scholte. Fantastic DNimals and where to find them. *Neuroimage*, 180(Pt A): 112–113, Oct. 2018.

H. S. Scholte, S. Ghebreab, L. Waldorp, A. W. M. Smeulders, and V. A. F. Lamme. Brain responses strongly correlate with weibull image statistics when processing natural images. *J. Vis.*, 9(4):29.1–15, Apr. 2009.

M. Schrimpf, J. Kubilius, M. J. Lee, N. A. Ratan Murty, R. Ajemian, and J. J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, Sept. 2020.

N. Seijdel, N. Tsakmakidis, E. H. F. de Haan, S. M. Bohte, and H. S. Scholte. Depth in convolutional neural networks solves scene segmentation. *PLoS Comput. Biol.*, 16(7): e1008022, July 2020.

K. A. Sundberg, J. F. Mitchell, T. J. Gawne, and J. H. R. Reynolds. Attention influences single unit and local field potential response latencies in visual cortical area V4. *J. Neurosci.*, 32(45):16040–16050, Nov. 2012.

S. Treue and J. C. Martínez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, June 1999.

R. Vallat. Pingouin: statistics in python. *JOSS*, 3(31):1026, Nov. 2018.

R. VanRullen. Perception science in the age of deep neural networks. *Front. Psychol.*, 8: 142, Feb. 2017.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, 17(3):261–272, Mar. 2020.

M. Waskom, O. Botvinnik, M. Gelbart, J. Ostblom, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. Warmenhoven, J. B. Cole, J. de Ruiter, J. Vanderplas, S. Hoyer, C. Pye, A. Miles, C. Swain, K. Meyer, M. Martin, P. Bachant, E. Quintero, G. Kunter, S. Villalba, Brian, C. Fitzgerald, C. G. Evans, M. L. Williams, D. O'Kane, T. Yarkoni, and T. Brunner. mwaskom/seaborn: v0.11.0 (sepetmber 2020), Sept. 2020.

V. Wyart, A. C. Nobre, and C. Summerfield. Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proc. Natl. Acad. Sci. U. S. A.*, 109(9):3593–3598, Feb. 2012.

D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.*, 19(3):356–365, Mar. 2016.

D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 111(23):8619–8624, June 2014.

Y. C. Yoon. LIF and simplified SRM neurons encode signals into spikes via a form of asynchronous pulse Sigma-Delta modulation. *IEEE Trans Neural Netw Learn Syst*, 28 (5):1192–1205, May 2017.

A. J. Yu and P. Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46(4): 681–692, May 2005.

D. Zambrano, R. Nusselder, H. S. Scholte, and S. M. Bohté. Sparse computation in adaptive spiking neural networks. *Front. Neurosci.*, 12:987, 2018.

Y. Zhang, E. M. Meyers, N. P. Bichot, T. Serre, T. A. Poggio, and R. Desimone. Object decoding with attention in inferior temporal cortex. *Proc. Natl. Acad. Sci. U. S. A.*, 108(21):8850–8855, May 2011.